

Towards a physical map of the *Drosophila melanogaster* genome: mapping of cosmid clones within defined genomic divisions

I.Sidén-Kiamos¹, R.D.C.Saunders², L.Spanos¹, T.Majerus³, J.Treanear³, C.Savakis¹, C.Louis¹, D.M.Glover², M.Ashburner³ and F.C.Kafatos^{1,4}

¹Institute of Molecular Biology and Biotechnology, Research Center of Crete, PO Box 1527, Heraklion 711 10, Crete, Greece, ²Cancer Research Campaign Laboratories, Department of Biochemistry, University of Dundee, Dundee DD1 4HN, ³Department of Genetics, Cambridge University, Downing Street, Cambridge CB2 3EH, UK and ⁴Department of Cellular and Developmental Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA

Received July 30, 1990; Revised and Accepted September 11, 1990

ABSTRACT

A physical map of the *D. melanogaster* genome is being constructed, in the form of overlapping cosmid clones that are assigned to specific polytene chromosome sites. A master library of ca. 20,000 cosmids is screened with probes that correspond to numbered chromosomal divisions (ca. 1% of the genome); these probes are prepared by microdissection and PCR-amplification of individual chromosomes. The 120 to 250 cosmids selected by each probe are fingerprinted by *HinfI* digestion and gel electrophoresis, and overlaps are detected by computer analysis of the fingerprints, permitting us to assemble sets of contiguous clones (contigs). Selected cosmids, both from contigs and unattached, are then localized by *in situ* hybridization to polytene chromosomes. Crosshybridization analysis using end probes links some contigs, and hybridization to previously cloned genes relates the physical to the genetic map. This approach has been used to construct a physical map of the 3.8 megabase DNA in the three distal divisions of the X chromosome. The map is represented by 181 canonical cosmids, of which 108 clones in contigs and 32 unattached clones have been mapped individually by *in situ* hybridization to chromosomes. Our current database of *in situ* hybridization results also includes the beginning of a physical map for the rest of the genome: 162 cosmids have been assigned by *in situ* hybridization to 129 chromosomal subdivisions elsewhere in the genome, representing 5 to 6 megabases of additional DNA.

INTRODUCTION

One of the earliest, and most profound, contributions of research with *Drosophila melanogaster* to biology was Sturtevant's discovery of genetic maps (1). Although the linkage of genetic factors had been found in sweet peas some years earlier (2), Sturtevant realized that the frequencies of recombinant phenotypes

could be used to obtain a linear order of genes. By 1925, the outlines of the linkage map of *D. melanogaster* were complete, with over 100 different genes ordered among four chromosomes (3). Generations of geneticists have improved and refined this map, and today approximately 3,800 different genes have been localized.

Since 1913 there have been three major contributions to the mapping of the *Drosophila* genome. The first of these followed Painter's discovery of giant polytene chromosomes in the larval salivary glands of *D. melanogaster* (4,5). The most remarkable feature of these chromosomes, apart from their size, is the constancy of their pattern of bands and interbands. Painter mapped this pattern and, by studying the chromosomes of mutant strains, demonstrated the colinearity of the cytogenetic and linkage maps. Painter's polytene chromosome maps were soon replaced by those of Calvin Bridges (6). Bridges divided each of the five major chromosome arms into 20 numbered divisions, and each division into six lettered subdivisions, thereby establishing an enduring system of reference. In this mapping system, each of the five major chromosome arms is divided into 20 divisions (*i.e.* the X is divisions 1–20, the left and right arms of chromosome 2 are 21–40 and 41–60 respectively, and those of chromosome 3 are 61–80 and 81–100 respectively). The very small 4th chromosome is assigned just two divisions, 101 and 102. Subsequently, when Bridges and his son revised his maps (7, 8, 9, 10, and 11), each band was given a number, so that it could be uniquely referenced by the combination of division number, subdivision letter and band number; *e.g.* 10B3 is the third band within subdivision B of division 10. The number of bands per lettered subdivision averages 8.6, with a range from 28 (in 66A) to 2 (in five subdivisions). Correspondingly there is a 10-fold range in DNA content per subdivision, from about 520 to about 50kb (12). Although the precise interpretation of the Bridges maps have been discussed by cytogeneticists (see, for example, 12,13), they remain an invaluable and, in their accuracy and detail, unique resource. The revised maps of Bridges and Bridges show 5059 bands, a number not very different from the

5076 estimated from electron micrographs (12). The polytene maps are tied to the genetic map in great detail, by means of numerous rearrangements (deletions, duplications, inversions and transpositions).

The second great advance in mapping the *Drosophila* genome followed from the development of techniques for the *in situ* localization of nucleic acids to polytene chromosomes (14). This allowed the precise assignment of purified nucleic acids to particular chromosome bands. With the third major advance, molecular cloning, over 2200 different genes have been located to individual bands, or at least to lettered subdivisions. In the early 1970s, Hogness' laboratory initiated the cloning of segments of *Drosophila* DNA with the intent of correlating physical and genetic maps, and so understanding the structure of bands and interbands in the salivary gland chromosomes. Initially, chromosomal DNA was cloned at random (15, 16), but a technique was soon developed permitting the selection of genes by a corresponding RNA probe (17,18,19). A strategy was then devised making it possible to 'walk' from one cloned sequence, mapped to the polytene chromosomes by *in situ* hybridization, towards a specific gene, by the isolation of overlapping segments of DNA cloned in phage vectors (20). Subsequently, additional cloning strategies have been developed, that of transposon tagging (21) being of continuing importance (*e.g.* 22, 23). Thus, in principle any *Drosophila* gene revealed by a mutation can now be cloned with reasonable effort.

Despite the relative ease of these approaches, a total physical map of *Drosophila* DNA would have enormous value. With freely available clones, it would result in major savings of effort for the entire community of fruitfly researchers, avoiding the duplication inherent in decentralized, piecemeal cloning strategies. A complete physical map would also be a necessary prelude to any project to sequence the *D. melanogaster* genome in its entirety. Beyond these considerations, a complete physical map would permit novel studies of chromosome organization and structure. It would, for example, revolutionize studies of chromosome evolution, and allow the analysis of presently poorly characterized relatives, including mosquitoes and other flies that are of major economic and medical importance to humans.

Complete physical maps are becoming available, for both procaryotic (24) and selected eucaryotic genomes: *Saccharomyces cerevisiae* (25,26), *Caenorhabditis elegans* (27,28) and *Arabidopsis thaliana* (29). Many groups are engaged in the construction of a map for the human genome. The eucaryotic maps consist of ordered clone libraries, more specifically groups of contiguous clones (*contigs*), whose DNA overlaps have been inferred from significant sharing of restriction enzyme fragments. The *Drosophila* genome is highly favourable for constructing a physical map, because of the existence of polytene chromosomes, the great wealth of genetic and molecular data now available, and its relatively small size (double that of *C. elegans* and *Arabidopsis*, but only one twentieth that of humans or other mammals).

Three major efforts to construct a physical map of the *Drosophila* genome are currently in progress. Hartl and collaborators are mapping large DNA segments, cloned in yeast artificial chromosomes (YACs), to the polytene chromosomes (30). J.D. Hoheisel and H. Lehrach (31) are assembling cosmid contigs selected by their hybridization to oligonucleotides of random sequence. We are focussing on the construction of a detailed map of cosmid contigs by restriction enzyme fingerprinting and *in situ* hybridization to polytene chromosomes.

Here we discuss the principles of our approach and its application to mapping the three terminal divisions of the X chromosome.

MATERIALS AND METHODS

Unless otherwise indicated all molecular procedures were performed as described in Sambrook *et al* (32).

Construction of the master library

Genomic DNA prepared from newly hatched adult Oregon-R (Cambridge) flies was partially digested with *Sau3A* and sized by agarose gel electrophoresis. Fragments corresponding to 30–50 kb in length were then isolated and subsequently ligated into the *Bam*HI site of the cosmid vector, Lorist 6 (33). Gigapack (Stratagene) packaging extracts were used to infect *E. coli* ED8767 cells according to the manufacturer's specifications. Individual colonies were then picked, gridded on Hybond-N (Amersham) filters and also stored in 25% glycerol at -80°C in microtitre plates. Each filter carries the equivalent of 8 microtitre plates. The total master library consists of 19,200 clones stored in 200 microtitre plates and gridded on 25 filters of 768 clones each.

Screening the master library and fingerprinting

The master library is screened repeatedly with DNA microdissected from a particular division and either cloned in phage λ NM1149 (34) or directly PCR-amplified as described elsewhere (35). To radioactively label the inserts of the microclones, the *Drosophila* DNA is first amplified by PCR using primers flanking the cloning site and then labelled by extension using the Klenow fragment of DNA polymerase I. In the case of directly microamplified material the probe is prepared by random priming (36). In both cases high specific activity (3000Ci/mmol) ^{32}P -dATP and ^{32}P -dCTP (Amersham) are used. Both pre-hybridizations and hybridizations are carried out at 65°C in $6\times\text{SSC}$, $5\times\text{Denhardt's}$, 0.5% SDS. Hybridization is allowed to proceed for 48 hours, and the filters are washed once in $2\times\text{SSC}$ (15 min) and twice in $2\times\text{SSC}$, 0.1% SDS (30 min) and then exposed for 72 hours using Royal X-OMat autoradiography film (Kodak). Fingerprinting of positive clones is carried out as described by Coulson and Sulston (28). Sets of *Hinf*I digests from eight cosmids are resolved by electrophoresis on adjacent lanes of a sequencing gel, alternating with lanes that carry size markers (λ DNA digested with *Hinf*I).

Fingerprint Analysis

For digital characterization and comparison of DNA fragments, we are using the system originally developed by Sulston and co-workers (27, 28, 37, 38), with minor modifications. The programs, whose code was kindly provided by J. Sulston, are written in VAX Fortran 77 and run on a microVAX II computer with the VMS operating system. A digital scanning densitometer (Autoreader, Amersham International) is used for semi-automatic data entry, and an AED 767 raster graphics workstation permits viewing and editing the data derived from the densitometer. A Selanar HIREZ 100XL terminal or a Macintosh Plus microcomputer running VT100/Tektronics 4014 terminal emulator software are used for graphic display and manual assembly of contigs.

Data Entry

Autoradiograms are scanned with the Autoreader at a resolution of 0.1 mm in the vertical axis and 0.4 mm in the horizontal axis. After transferring to the VAX, the digitized image of the

autoradiogram is processed by computer. This step involves automatic recognition of the lanes containing size markers and those containing fingerprints, identification of individual bands, and calculation of the optimal alignment between detected and expected size markers. The autoradiogram is then displayed on the AED and artefactual bands are removed by the operator; little or no such editing is required with most of our autoradiograms. The distances, from the origin of the gel, of the accepted bands of each fingerprint are then normalized by the computer against an extrapolation from the flanking markers, and are added to the appropriate database.

Data Storage and Contig Assembly

The data from each division are stored in a separate database, which consists of a set of direct access files containing the following information for each clone: clone name, normalized band coordinates, contig number to which the clone belongs, position of the clone within the contig, and additional remarks, such as position in the polytene chromosomes and genes encompassed. To determine possible overlaps between clones, all clone pairs are compared, and for each pair the number of bands that match within a preset tolerance (1.0 mm) is recorded. The probability that the observed number of band matches in a pair of clones is due to chance (*i.e.* the probability that the two clones are not really overlapping) is then calculated from the total number of bands in each clone and the tolerance. All matches for each clone to other clones in the database are ranked in terms of this probability, and those better than an operator-set cutoff are printed out.

Initially, we used the printed output to assemble contigs semi-automatically. We found, however, that an automatic contig assembly algorithm described by Sulston *et al.* (37) gave very satisfactory results, and we are currently using this for contig assembly. The program (routine CONTASP of the Sulston programs) first performs a preliminary sorting of the entire database, in which contigs are defined as groups of related clones. Within each group, every member can be linked to others by chains of matches in which no link has a probability of coincidence (*i.e.* probability that the event is due to chance) higher than a preset value (10^{-5} for the first analysis, see Results). After the initial assembly, the program positions all clones within each contig relatively to each other, beginning with the pair having the highest number of matches and adding successively clones with the next highest match to any of the already positioned clones. As each clone is introduced, all possible positions are considered, and a position is chosen as described by Sulston *et al.* (37). After editing of the database, final refinement of the contigs is done by taking into account hybridization data, as described in the Results.

In situ hybridization

Probes were labelled with Bio-16-UTP (BCL-Boehringer Mannheim) either by nick-translation or by priming with random oligonucleotides. The probes were hybridized to the polytene chromosomes of *D. melanogaster* (Canton-S) or *D. simulans* (C167.4, a wild-type stock collected in Kenya, October 1973), which had been alkali denatured after squashing (39, 40). Hybridization was detected using streptavidin-horseradish peroxidase (ENZO Biochemicals) and, after color development with 3,3'-diaminobenzidine, the chromosomes were stained with Giesma at pH 6.8. The preparations were mounted in Depex and then photographed on color negative film under phase contrast using a Zeiss Axiophot microscope. The polytene chromosomes

were interpreted using the revised maps of Bridges and Bridges (13), and *in situ* positions are given as limits, unless a cosmid spans more than one identifiable site when the distal and proximal limits are linked by the work 'to'.

RESULTS

General strategy

In a large-scale genomic mapping project, the first step is to construct a master library which is subsequently ordered. We have made such a library in the Lorst 6 cosmid vector (33). It contains nearly 20,000 clones, individually picked and stored in single tubes. For purposes of hybridization the library is also represented as sets of 25 filters, each filter spotted with 768 clones in a densely gridded pattern. A conservative estimate of the insert size is 35kb. Therefore, assuming that the *Drosophila* genome is 165Mb (41), this library represents more than a 4-fold coverage of the genome. Our strategy for ordering the library is diagrammed in Fig.1.

J. Sulston and his collaborators pioneered a general approach for assembling sets of overlapping cosmid clones or *contigs*, by restriction enzyme finger-printing followed by computer analysis (27,28,37). We have used the same approach but with important modifications that take advantage of a highly favourable feature of *Drosophila*, its polytene chromosomes. Instead of fingerprinting random cosmids and assembling them into contigs that are scattered throughout the genome, we use hybridization to preselect sets of neighbouring clones, and thus assemble the contigs in order, beginning with the tip of the first chromosome, the X. The probes for this hybridization are derived from DNA of individual chromosomal divisions (as defined by Bridges (7), see also Lefevre (13)), *i.e.* each one should represent approximately 1% of the genome. Indeed, we find that, when used to screen the filters of the library, each division-specific probe selects 120 to 250 cosmids. The great advantage of this

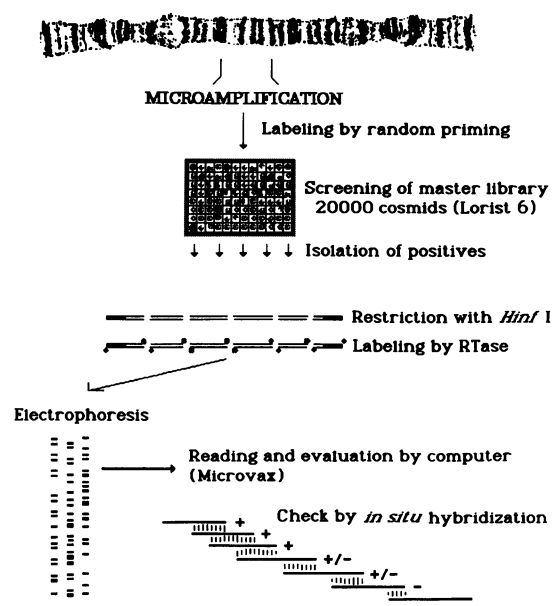


Fig. 1. Outline of our physical mapping strategy, from microamplification of chromosomal divisions (top), through screening a master cosmid library to fingerprinting and assembling into contigs the selected clones, which are finally checked by hybridization to polytene chromosomes. For details see Text.

approach is that, at each step, we are analyzing only one out of 102 small 'genomes', with an average size of 1.2×10^6 bp, instead of the entire *Drosophila* genome of 165×10^6 bp¹. The number of all pairwise combinations of clones, which must be examined to establish overlaps, is nearly proportional to the square of the number of clones being analyzed. Thus when 100 sets of neighbouring clones are used, each set corresponding to 1% of the genome, the total number of false overlaps will be approximately 100-fold lower than if the entire library were analyzed simultaneously at the same criterion of stringency. Effectively, at each step a much higher proportion of the fingerprinted clones are assembled into contigs by our approach than if we were dealing with clones from throughout the genome (See Materials and Methods).

After contig assembly, representative clones are mapped by *in situ* hybridization to polytene chromosomes, to correlate then with the cytogenetic map and to confirm the validity of fingerprint analysis (consistency of chromosomal localization of the clones comprising each contig). Terminal cosmids from the contigs also are used to determine the pattern of cross-hybridizations, thereby linking some contigs or adding additional clones to them. Finally, hybridizations are performed with probes from genes that have been previously cloned by others, and mapped within the respective division. Taking all data into account, a canonical set of cosmids is then selected to represent the physical map of the chromosomal division, which is anchored to both the genetic and the cytogenetic maps.

Division-specific probes and hybridizations

The probes are generated by microdissection of a single division from a single polytene chromosome, followed by PCR-microamplification (35,44,45,46, see also 47). Initially, we experimented with microcloning the dissected DNA into the insertion vector NM1149 (34, 48). However, in our hands that procedure proved relatively unreliable, often yielding probes of low complexity (*i.e.* a low proportion of microdissected DNA was represented in clones; (35)), and was abandoned after constructing the division 3 map. In the microamplification procedure used for subsequent work, the microdissected DNA is digested with restriction endonuclease *Sau3A*, double-stranded oligonucleotides are ligated to the ends, and the fragments are amplified by the PCR technique, using primers that correspond to the oligonucleotides. Effectively, each microdissected segment is perpetuated as a pooled division-specific probe.

In preliminary experiments, in which DNA was microdissected from *D. melanogaster* chromosomes and used to probe the *D. melanogaster* master library, hybridization with dispersed repetitive DNA proved to be a considerable problem, especially in certain divisions. About 20% of the *D. melanogaster* genome is estimated to be middle repetitive DNA (49). To circumvent this problem, we now prepare the division-specific probes from *D. simulans*. In this sibling species only 3% or so of the genome is middle repetitive (49), and few of its repetitive sequences are also found in *D. melanogaster* (50), at least in high copy number (51). The nucleotide divergence between *D. melanogaster* and *D. simulans* is quite limited: the majority of their single-copy sequences are divergent to the extent of only 3–4% substitution, although a fraction is substantially more divergent (perhaps

12–17% substitution for approximately 20% of the genome) and does not crosshybridize under conditions more stringent than our (52, 53). Importantly, the chromosomes of the two species are essentially homosequential, *i.e.* show the same polytene chromosome banding pattern with the exception of a major inversion on 3R and a few very small inversions (54). Since our probes are relatively depleted of sequences that are repetitive in *D. melanogaster*, they can be used to select sets of division-specific *D. melanogaster* cosmids without gross contamination with extraneous clones sharing repetitive DNA.

The advantage of using *D. simulans* is documented in Fig. 2. In the top panel, *D. melanogaster* chromosomes are hybridized with probes derived from division 10, either from *D. melanogaster* (Fig 2A) or from *D. simulans* (Fig. 2B). The noise level of probe hybridization to other chromosomal regions is evidently higher for the homospecific probe. The bottom panel exemplifies an additional use of *D. simulans* when a *D. melanogaster* cosmid that contains substantial sequence repeats is to be analyzed, *in situ* hybridization to *D. simulans* chromosomes often permits accurate mapping (Fig 2D), whereas homospecific hybridization is less definitive (Fig. 2C).

The quality of our probes is illustrated in Fig. 3. The distal part of an X chromosome is presented, which has been hybridized with the probe corresponding to division 2. The virtual absence of hybridization except in the pertinent division is noteworthy. A similar low background is evident over the rest of the chromosomes (data not shown). Fig. 3 also includes X chromosome segments that hybridize with an additional ten division-specific probes, each one aligned with the division of probe origin. It is clear that the probes are of high specificity (do not hybridize strongly outside the pertinent division), and of high complexity (hybridize throughout the division, in a pattern that roughly parallels the inherent intensity of the bands, *i.e.* their DNA content).

In screening the master library (Fig. 4), we experience some loss of specificity and complexity. Although strong positives are unambiguous, clones of interest can be missed because of poor growth of bacterial colonies, and false positives can be obtained by hybridization with residual repeats in the probes. As we show below, the data corruption from false positives is not serious, since our canonical set mostly consists of cosmids whose localization has been confirmed by *in situ* hybridization to polytene chromosomes. Furthermore, there is no substantial loss of effort. An individual edited database has been set up for each chromosomal division, encompassing all cosmids which have been shown by *in situ* hybridization to belong to that division, plus any cosmids that have been linked to them in contigs. False positives revealed by *in situ* hybridization, and randomly mapped clones from preliminary experiments, are assigned to the appropriate edited database and will be analyzed there together with their real neighbours. Thus far we have characterized by *in situ* hybridization a total of 470 cosmids, of which 305 have been definitively assigned to a chromosomal site (see below). Approximately half of these mapped clones are from the three distal divisions of the X chromosome, but the rest represent 129 additional subdivisions, *i.e.* 23% of the total for divisions 4–102.

Cosmid fingerprinting and contig assembly

Cosmids selected with division-specific probes were fingerprinted using complete *HinfI* digests (55); in our hands the *HindIII/Sau3A* protocol used to generate labelled DNA fragments of *C. elegans* (27) was not fully reliable, because of frequent partial digestion of cosmids. The *HinfI* fragments were end-labelled with reverse

¹ Although the *Drosophila* genome is estimated as 165 Mb, allowance must be made for the heterochromatin (notably its large centromeric fraction), which is underreplicated in polytene chromosomes (42, 43). Thus, we expect, that the euchromatin of the chromosomal arms contains ca 75% of the genome, or 1.2 Mb per division, on the average.

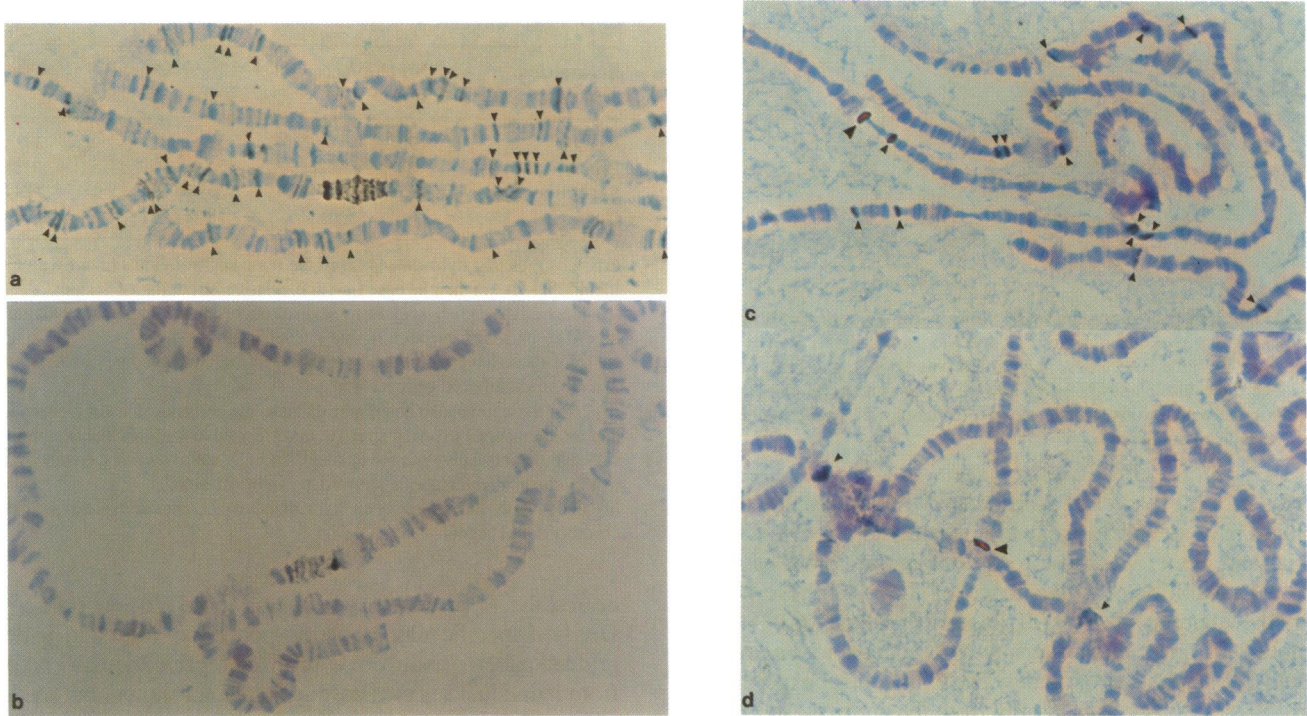


Fig. 2 (A,B). A comparison of the hybridization to the polytene chromosomes of *D. melanogaster* to PCR amplified material from division 10 of a X chromosome of *D. melanogaster* (A) or from the same division from an X chromosome of *D. simulans* (B). Both probes label division 10, as expected, but the homospecific probe also labels a large number of other sites dispersed throughout the genome; some of these are marked by arrows in (A). (C,D) An example of a cosmid (190A11) which includes a middle repetitive sequence. In *D. melanogaster* (C), 21 different sites are labelled in the chromosome arms (4 on the X, 3 on 2L, 3 of 2R, 6 on 3L and 5 on 3R; arrows). Despite this, the site from which the clone originated can be determined as being 47A (large arrow), since this is the most intensively labelled site. This is confirmed by *in situ* hybridizing the same probe to the chromosomes of *D. simulans* (D): the 47A site is very heavily labelled and there are only two other weak sites of hybridization (small arrows).

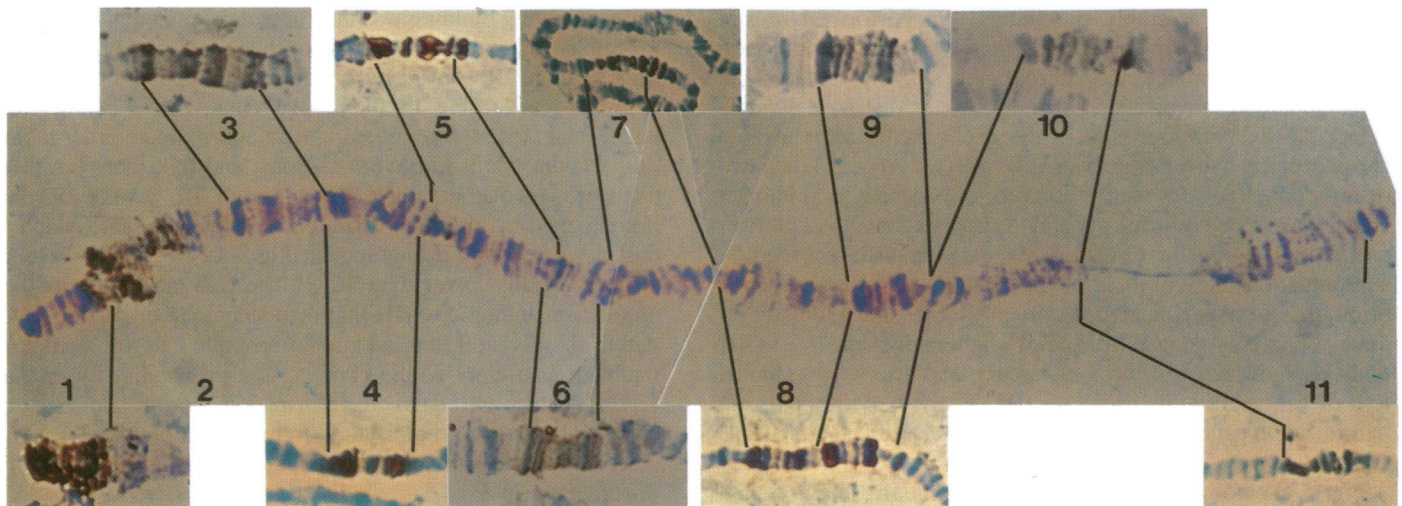


Fig. 3. *In situ* hybridization of X chromosomes of *D. melanogaster* with PCR amplified DNA microdissected from specific divisions of the X chromosome of *D. simulans*. In the center is a distal X chromosome (divisions 1 to 11), hybridized with the microamplified probe for division 2. Alongside this chromosome, both above and below, are segments showing the results of *in situ* hybridizations with ten additional division specific probes, as indicated.

transcriptase and analyzed directly by electrophoresis, followed by autoradiography, densitometry and computer-aided detection of coincident bands (38). *HinfI* yielded an average of 60 detectable bands per clone. For each pair of cosmids the number of bands that comigrated within 1mm tolerance was recorded, and the

probability of random coincidence (*i.e.* the probability that this number of matches is due to chance) was calculated as a function of the number of bands in each cosmid and the tolerance (37). Because of the small size of each 'genome' being analyzed (*ca.* 1.2 Mb and 200 cosmids per division), it was possible to detect

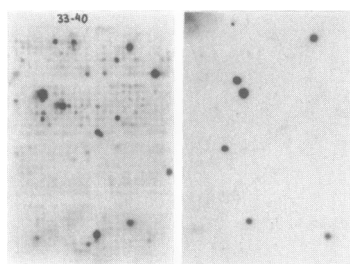


Fig. 4. Examples of screening the master library. Two filters are shown, from experiments using two different division-specific probes. Each filter is densely gridded with the same 768 clones, in a pattern corresponding to eight microtitre dishes. The filter on the left (33–40) shows a background of all clones, as well as approximately 14 positives. The other filter has lower background, and its 6 positives are unambiguous.

overlaps at a high level of confidence without demanding that the clones have an unreasonably high number of coincident bands. By trial and error, we chose to do our initial contig assembly using a probability of coincidence of 10^{-5} as the criterion. Given the average number of bands per clone (60) and the tolerance (1 mm), this criterion corresponds to an average nominal overlap of approximately 50%. In control experiments, when 450 clones picked randomly from the entire master library were analyzed in three sets of 150, only 5 out of 33,525 possible cosmid pairs were scored as overlapping at this probability criterion. This result was in good agreement with the expected number of true contigs, given the size of the *Drosophila* genome, the number of clone pairs analyzed, and the minimum detectable overlap (56). In a second round of contig assembly, we used the statistically very stringent criterion of ten-fold lower probability (10^{-6} , see below).

In situ hybridizations and refinement of the map

After the initial assembly of contigs, we performed *in situ* hybridization analysis on cosmids putatively located at contig ends, as well as on selected internal clones of long contigs and some apparently unattached cosmids. Clones that gave a single detectable hybridization site in the euchromatic arms and no chromocentric hybridization were classified as unique. When two or more labelled sites were observed, any clearly predominant signal was scored as primary, and the approximate total number and some locations of secondary signals were also recorded. A substantial database of *in situ* hybridization data was assembled through these experiments, comprising 470 cosmids to date. Of these, 143 were assigned to a site of origin in divisions 1–3 (102 uniquely hybridized to these divisions), and 162 were assigned to a site of origin in divisions 4–102. The variety of the *in situ* hybridization patterns is illustrated by their classification in Table 1.

Fig. 5 exemplifies the *in situ* results. Data are shown for 23 cosmids, which are clearly localized at distinct sites, from distal to proximal across divisions 1, 2 and 3. An example is also shown of a repetitive clone that shows no primary hybridization site.

To refine the contigs, and to begin linking contigs and anchoring them to the genetic map, selected cosmids putatively located at contig ends were used for crosshybridization analysis. A convenient feature of the Loris 6 vector is that it has promoters for SP6 and T7 RNA polymerases, flanking and directed towards the foreign DNA insert. From each selected cosmid, pooled end probes were made by using both SP6 and T7 RNA polymerases. In practice, the probes predominantly represented sequences

Table 1. Patterns of *in situ* hybridization of cosmids to the polytene chromosomes of *D. melanogaster*

Euchromatic sites	No chromocentral signal	Chromocentral signal	Total
Unique	190	12	202
Primary (a)	68	34	102
Repetitive (b)	25 (c)	123 (d)	148
None	—	18 (e)	18

(a) Cosmids that hybridize to more than one euchromatic site but with a clearly identifiable site of origin based on the relative strengths of signals.

(b) Cosmids which hybridize to several euchromatic sites (the maximum was 80–100) and with no identifiable site of origin.

(c) 3 of these have also been hybridized to the chromosomes of *D. simulans*: 2 had an identifiable primary site.

(d) 26 of these have also been hybridized to the chromosomes of *D. simulans*: 9 had an identifiable primary site, 10 were repetitive with no identifiable primary site, (although usually with fewer sites than in *D. melanogaster*), 5 only hybridized to the chromocenter and 2 showed no hybridization.

(e) 4 of these have also been hybridized to *D. simulans* and showed only chromocentral signal.

nearest the T7 promotor, which gave more efficient transcription. Division-specific clones were spotted on filters in a densely gridded pattern, and were hybridized repeatedly with end probes from individual selected cosmids (data not shown, but see legend of Fig. 6). For quality control of the contigs, a possible alternative to crosshybridization is fingerprinting with a second enzyme; this alternative was tested for division 3, using *Ban*I digestion, and proved satisfactory but relatively time consuming. The division-specific filters were also screened with probes for selected genes which have been cloned and characterized by others.

Taking into account the *in situ* hybridization data, crosshybridizations between cosmids, fingerprints with a second enzyme, and hybridizations between cosmids and known genes, an edited database was constructed for each division, and a second round of contig building was then undertaken. Contigs were first assembled by computer analysis at the more stringent criterion of less than 10^{-6} probability, and were linked or modified manually, according to the above data. First, contigs and unattached clones were placed into an approximate order taking into account the *in situ* hybridizations, and links between them that were less than definitive ($10^{-3} > p > 10^{-6}$) were accepted if confirmed by crosshybridization. Clones were also added to the ends of contigs if they matched at least two internal clones at a probability lower than 10^{-5} . The canonical set was then defined, including cosmids that map to unique or primary sites within the division, and additional cosmids that are needed to maintain continuity within contigs, at a probability lower than 10^{-6} . A list and detailed information on *in situ* mapped canonical clones of the distal X divisions are presented in Fig. 2; cosmids that belong to contigs are omitted, as they are fully documented in Fig. 6.

In the computer-assisted assembly of contigs, order is determined by the degree to which a cosmid matches all of its neighbouring clones (see Materials and Methods), and is subject to some uncertainty. For final contig refinement, minor changes in the order of clones were incorporated, to provide for consistency with the *in situ* and crosshybridization data. Neighbouring clones within a contig were aligned so that their apparent overlap is proportional to their number of co-migrating bands. No effort was made to control the alignment of clones that are not nearest neighbours. Neighbours may appear to have a greater overlap than in fact is the case, because of random

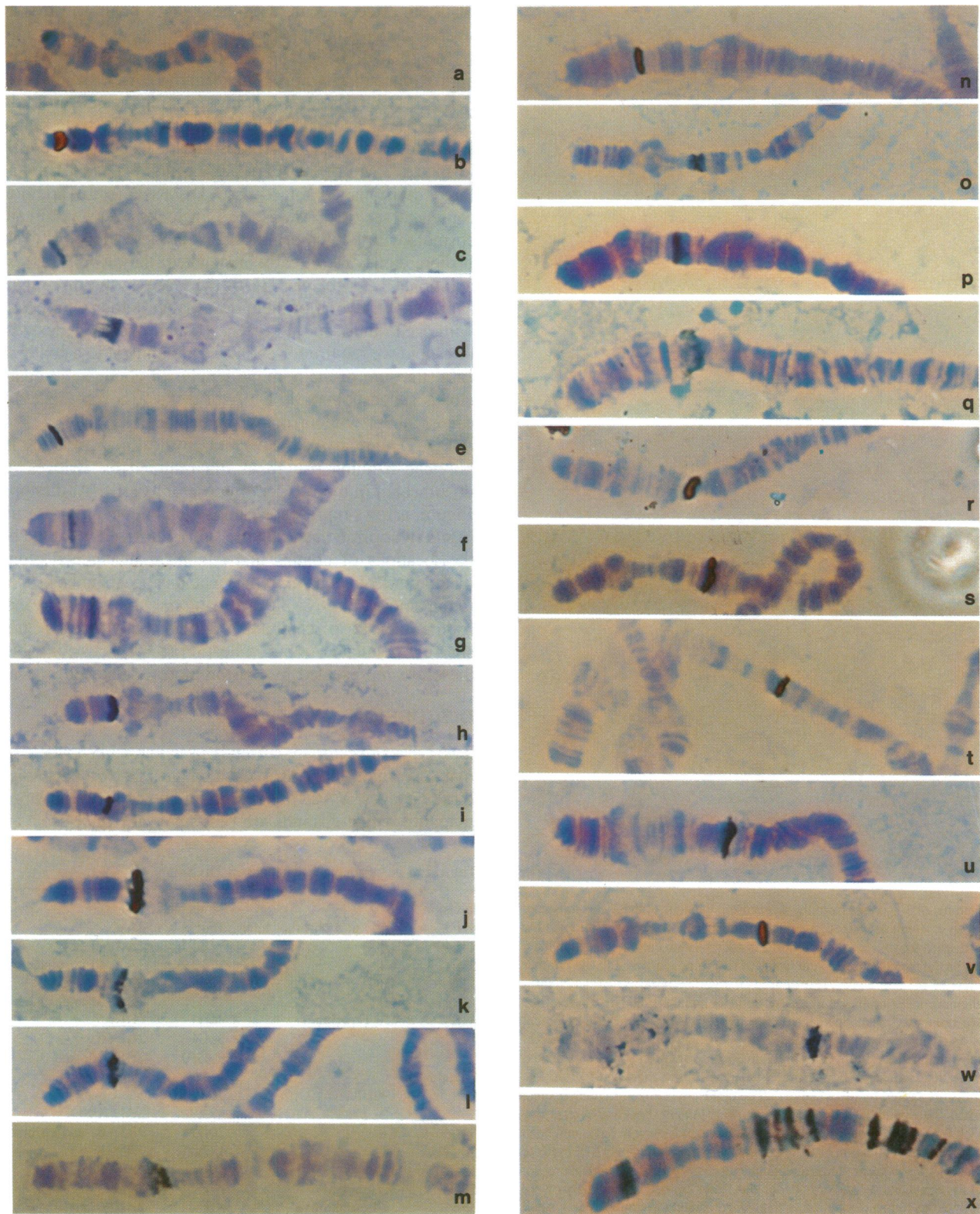


Fig. 5. Examples of *in situ* hybridizations with cosmid clones. Panels *a* to *w* show the primary sites of hybridization of selected cosmids to region 1–3 of the X chromosome of *D. melanogaster*. The clones and their site of origin are as follows. Clone (*a*) 23E12/1A1.2; (*b*) 180C11/1B1–4; (*c*) 16C9/1B3–10; (*d*) 115C2/1B10–14; (*e*) 11F6/1B; (*f*) 133C8/1D1.2; (*g*) 35F1/1E; (*h*) 8D8/1F; (*i*) 52C10/2A; (*j*) 70B2/2B1–4; (*k*) 17E4/2B4–7; (*l*) 153C9/2B11–18; (*m*) 129E12/2B9 to 2B16; (*n*) 38H12/2B10–18; (*o*) 62C3/2E to 2F; (*p*) 25E8/2F; (*q*) 94D5/3A; (*r*) 60G6/3C1–3; (*s*) 140G11/3C1–2; (*t*) 38B10/3C; (*u*) 114E2/3D; (*v*) 152F11/3E; (*w*) 170B5/3EF. Panel *x* gives an example of a cosmid (170G10) that shows a large number of sites on the distal X chromosome (and elsewhere). In fact, this precise pattern of repetitive sites is quite common; presumably 170G10 includes sequences of one of the common transposable elements of *D. melanogaster*.

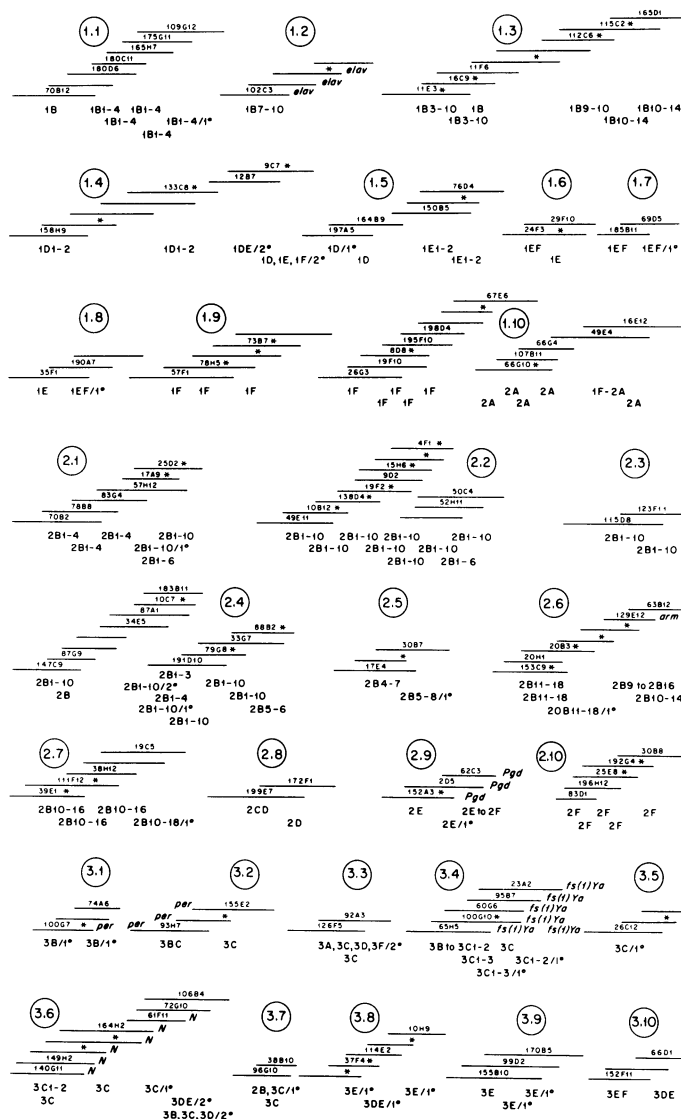


Fig. 6. Diagrams of 30 contigs from the three distal divisions of the *X* chromosome. Contigs 1.1 to 3.10 are presented in approximately distal to proximal order. One contig, 1.10, spans the division 1–division 2 border. Within each contig, individual cosmids are represented by horizontal lines, of length proportional to the number of bands scored by computer. The degree of overlap between adjacent lines indicates the number of comigrating restriction fragments of the two clones. The cosmids are numbered if mapped by *in situ* hybridization, and the cytological location is listed underneath, aligned with the left end of the cosmid number. Locations are mapped to subdivision and sometimes to within a range of bands, with uncertainties indicated. For cosmids that span more than one identifiable site, the limits are linked by the word 'to'. The clones hybridized to unique sites, unless marked 1° (repetitive, with one or more predominant sites) or 2° (repetitive without a clearly predominant site). Canonical cosmids which have not been mapped by *in situ* hybridization are shown without a number. Asterisks indicate the existence of one or more additional fingerprinted cosmids, which were not included in the canonical set and are not shown. Links between cosmids that were established by cross-hybridization rather than fingerprinting are indicated by small overlaps: the links between clones 133C8 and 12B7 (contig 1.4), 164B9 and 150B5 (contig 3.6), and 66G4 and 49E4 (contig 1.10). Cosmids 164H2 and 61F11 (contig 3.6) have been linked by hybridization to a *Notch* probe. Cosmid hybridization to probes of genes cloned by others is indicated by italicized symbols (*elav*, *arm*, *Pgd*, *per*, *fs(1)Ya*, *N*; kindly provided by K. White, M. Peifer, E. Wieschaus, J.-M. Dura, M. Young, M. Wolfner and S. Artavanis-Tsakonas respectively).

coincidence of bands in the fingerprints; or they may appear to be less overlapping than in fact is the case, because of failure to detect coincident bands. The cross-hybridization data suggest

that both effects occur, with the balance being towards a slight overall underestimate of overlaps.

The final division 1–3 contigs (Fig. 6) were numbered in approximate distal to proximal order (1.1 ... 1.10, 2.1 ... 2.10, 3.1 ... 3.10). These 30 contigs encompass 220 cosmids; 149 cosmids are canonical, and of these 108 have been mapped by *in situ* hybridization to chromosomes. The contigs are estimated to span 2.7 Mb of DNA, or 71% of the expected length of the three divisions combined (see Discussion). In addition, the canonical set of divisions 1–3 includes 32 as yet unattached cosmids with known primary sites in these divisions (Table 2). If these clones were really unattached, they would represent an additional 1.2 Mb, or 31% of the DNA in the three divisions. More likely several of them do overlap, either with a contig or with each other, but the extent of such overlap is less than the threshold we require for definitive assignment to a contig. In any case, the sum of the nominal lengths of contigs and unattached clones suggests that the DNA of distal *X* is represented in our canonical set without many major gaps.

Fig. 7 summarizes the current physical map, relating it to the polytene banding pattern as diagrammed by Bridges (6) and documented in more recent micrographs (13). At the top of the Figure, arrows represent cosmids with known unique or primary *in situ* hybridization sites in divisions 1 through 3. At the bottom, the locations of contigs and canonical unattached clones are diagrammed. In general, representation appears to be proportional to the amount of DNA in each band or subdivision, as indicated by the intensity of band staining (12). Only a few regions are sparsely populated in our collection (see Discussion), but even these are represented by one or more unattached cosmids.

DISCUSSION

Two types of vector are currently used to establish whole genome maps, cosmids and YACs; each one has its own advantages and disadvantages. YACs are vectors of choice for rapid large-scale mapping, while cosmids are best suited for finer-scale analysis

Table 2. List of unattached clones in Division 1–3*

Division 1		Division 2	
42F3	1A1, and chromocenter	154H3	2A
23E12	1A1	52C10	2A, and 58F
65F1	1B	123B12	2A, and chromocenter
122H9	1BC	4E6	2B1–4
118E1	1D	122E11	2B1–5, and 19CD
63D2	1E, repetitive	36D1	2B1–5
		36E4	2B1–5
		96G10	2B1–10, and 3C
Division 3			
	199H8	2B1–10	
		17A9	2B1–10
94D5	3A	20D1	2B3–8, and chromocenter
20F11	3C	112C5	2B5–8
131D1	3C	79H6	2B7–8
58G5	3D, and 8E	67C7	2B10–18
85E1	3DE, repetitive	81D4	2B10–18
155B5	3E, repetitive	133E12	2C
		67A9	2C1–4
		22E5	2C4–10
		103B4	2EF
		62D9	2F

*Canonical cosmids which have not been assigned to contigs and have unique or primary hybridization sites in the three distal divisions of the polytene *X* chromosome. Any additional hybridization sites are indicated, as is hybridization to the chromocenter. Three clones include repetitive sequences that hybridize to multiple secondary sites.

and are much easier to use. In essence, YAC maps may be likened to pages of an atlas, while cosmid maps are analogous to city street maps. We have opted to assemble a detailed map in cosmids, while Garza *et al.* (30) are mapping a collection of YACs to the polytene chromosome. When the two approaches converge, the *Drosophila* scientific community will have available an additional potent tool for molecular genetic analysis.

It is crucial for users to understand the significance and limitations of contig analysis. Contigs are *not* equivalent to chromosomal walks of the type pioneered by Bender *et al.* (20). A contig is a best-fit representation of similarities among a set of clones based on restriction fragment fingerprints, while the familiar chromosomal walk is a map of clones whose overlaps have been defined precisely by restriction enzyme mapping. As explained in the Results, our diagrams of contigs (Fig. 6) include nominal overlaps, representing the number of co-migrating bands in the fingerprints of neighbouring clones. The order of cosmids within a contig is also a best estimate, and in some cases a slightly different order is also tenable. We have incorporated *in situ* hybridization and cross-hybridization analysis as important quality control and refinements of contig assembly. The resulting contigs are sufficient to indicate the approximate location of a gene of interest, between known genetic or cytogenetic markers. While defining more precisely the DNA location of their target gene, future users undoubtedly will obtain restriction maps, and thus will convert contigs into chromosomal walks, refining the physical map that is available to the entire community.

The nominal length of DNA represented in contigs plus canonical unattached clones is estimated as 3.9 Mb, or 103% of the 3.8 Mb total DNA in divisions 1, 2 and 3 (approximately 1.2, 1.3, and 1.4 Mb, respectively (12, *cf* 57)). Our guess is that the representation may be overestimated by approximately

20%, although it is difficult to obtain upper and lower limits, since we do not know precisely the extent of clone overlap, nor the degree of undetected overlap between contigs and putatively separate cosmids (see Results). In the *C. elegans* mapping project, when putatively unattached contigs were joined by YAC clones, it became evident that gaps between contigs were usually very small, or even absent (57). Examination of the top diagram in Fig. 7 indicates that the most serious deficiency is in subdivision 3A, which is estimated to total 313 kb (12), but is represented by only one unattached canonical cosmid. This may be due to conservative microdissection in preparing the division 3-specific probe. Furthermore, the paucity of definitive division 3A cosmids, like the slight underrepresentation of the entire division 3 in our contig map (Fig. 6), may be explained by two additional considerations. The division 3 probe was constructed by microcloning rather than microamplification, and therefore was less representative; and many of the cosmids selected by this probe (which was from *D. melanogaster*) were omitted from the edited database, because they were dominated by repetitive DNA.

A by-product of our work to date is a second set of cosmid clones, which have been mapped to 129 chromosomal subdivisions throughout the genome (other than divisions 1, 2, and 3). Making allowance for probable overlaps between a few of these clones, we estimate that this second set represents 5–6 Mb of mapped DNA or 4–5% of the euchromatic genome of *Drosophila*. Many of these clones have been mapped to a few bands (data available on request), and should be valuable even at this stage for isolating genes of interest, *e.g.* by limited chromosomal walking. The continuation of this work to provide a detailed physical map of the rest of the X chromosome is well in hand, and will eventually be extended to the rest of the genome.

In parallel with its extension, the cosmid map should be

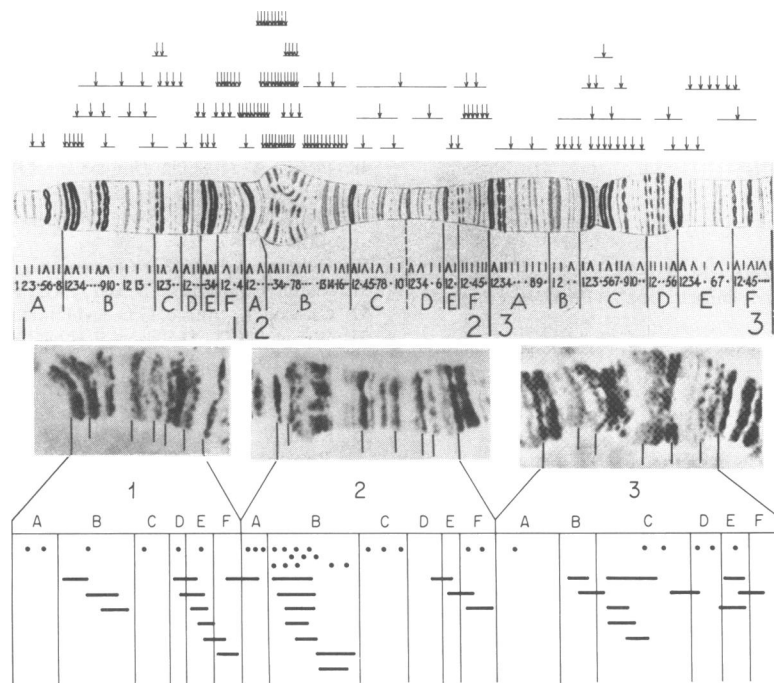


Fig. 7. A summary of the physical map of the three distal X divisions. Diagrams (7) and photographs (12) of these chromosomal regions are included. At the top, individual arrows indicate cosmids with *in situ* mapped unique or primary sites; horizontal lines, which refer to the Bridges diagram, indicate the accuracy of map assignment. At the bottom, the 18 subdivisions of this region (1A to 3F) are shown on a scale proportional to their estimated DNA content (12). Unattached canonical cosmids are shown as dots, and contigs as horizontal bars. The position and length of each bar reflect the contig location and its uncertainty, rather than overlap with adjacent contigs.

anchored in greater detail to the genetic map, by additional hybridization tests with genes cloned by others, and by *in situ* hybridizations to polytene chromosomes that have deletions, inversions or translocations. It will also be important to correlate the emerging cosmid and YAC maps. YACs will be especially valuable for establishing links, proximal-distal orientations and left-right ordering of contigs that have not been resolved cytologically.

An important rationale for this project is to serve the scientific community. The canonical cosmids defining the physical map will be freely available, beginning in November 1990, after the logistics of distribution have been addressed. Those interested should contact M. Ashburner (E-mail, MA11@phx.cam.ac.uk, FAX 44-223-333992). In the longer term, as the detailed physical maps become refined, we believe it would be worthwhile to determine short sequences from the ends of selected cosmids, thus defining sequence-tagged sites (STS) and obviating the need for distribution of clones (59). The ordered cosmid library, or its STS database derivative, will be a convenient starting point not only for studies in molecular genetics but also for global sequencing of the *D. melanogaster* genome.

ACKNOWLEDGEMENTS

We are grateful to J. Sulston and A. Coulson for their assistance in launching this project, for making available the computer programs for fingerprint analysis, and for helpful advice. We thank H. Jaekle, K. Wharton, and M. Monastirioti for participation in early phases of the project, P. Little for advice, G. Papagiannakis, V. Schoinas and V. Triantafyllou for technical assistance, M. Pittarokilis for assistance with computer analysis and all member of the IMBB Insect Group for helping 'toothpick' the master library. We are grateful to A. Gruner-Schlumberger for making possible the initiation of the project with seed funds from Foundation Schlumberger, and to the European Economic Community (Stimulation Action Programme) for continued financial support. We also acknowledge support from the Greek Secretariat General for Research and Technology, and from MRC and CRC grants to DG and RS, which enabled their participation in the earlier phases of this work.

REFERENCES

1. Sturtevant, A.H. (1913) *J. Exp. Zool.*, **14**, 43–59.
2. Bateson, W., Saunders, E.R. and Punnett, R.C. (1905) *Repts. Evol. Committee Royal Society* No 2.
3. Morgan, T.H., Bridges, C.B. and Sturtevant, A.H. (1925) *Bibliographica Genetica*, **2**, 1–262.
4. Painter, T.S. (1933) *Science*, **78**, 585–586.
5. Painter, T.S. (1934) *Genetics*, **19**, 175–188.
6. Bridges, C.B. (1935) *J. Heredity*, **26**, 60–64.
7. Bridges, C.B. (1938) *J. Heredity*, **29**, 11–13.
8. Bridges, C.B. and Bridges, P.N. (1941) *J. Heredity*, **30**, 475–476.
9. Bridges, P.N. (1941) *J. Heredity*, **32**, 64–65.
10. Bridges, P.N. (1941) *J. Heredity*, **32**, 299–300.
11. Bridges, P.N. (1942) *J. Heredity*, **33**, 403–408.
12. Sorsa, V. (1988), *Chromosome maps of Drosophila*. 2 vols. CRC Press, Boca Raton, Fla.
13. Lefevre, G., Jnr. (1976) In: *Genetics and Biology of Drosophila*. Vol 1A (eds) M. Ashburner and E. Novitski. Academic Press, London. pp.31–66.
14. Pardue, M.L., Gerbi, S.A., Eckhardt, R.A. and Gall, J.G. (1969) *Chromosoma*, **29**, 268–290.
15. Wensink, P.C., Finnegan, D.J., Donelson, J.E. and Hogness, D.S. (1974) *Cell*, **3**, 315–325.
16. Glover, D.M., White, R.L., Finnegan, D.J. and Hogness, D.S. (1975) *Cell*, **5**, 149–157.
17. Grunstein, M. and Hogness, D.S. (1975) *Proc. Natl. Acad. Sci., USA*, **72**, 3961–3965.
18. Artavanis-Tsakonas, S., Shedl, P., Tschudi, C., Pirrotta, V., Steward, R. and Gehring, W.J. (1977) *Cell*, **12**, 1057–1067.
19. Lis, J., Prestidge, L. and Hogness, D.S. (1978) *Cell*, **14**, 901–919.
20. Bender, W., Spierer, P. and Hogness, D.S. (1983) *J. Mol. Biol.*, **168**, 17–33.
21. Bingham, P.M., Levis, R. and Rubin, G.M. (1981) *Cell*, **25**, 693–704.
22. Cooley, L., Kelly, R. and Spradling, A. (1988) *Science*, **239**, 1121–1128.
23. O'Kane, C.J. and Gehring, W.J. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 9123–9127.
24. Kohara, Y., Akiyama, K. and Isono, K. (1987) *Cell*, **50**, 495–508.
25. Olson, M.V., Hutchik, J.E., Graham, M.Y., Brodeur, G.M., Helms, C., Frank, M., MacCollin, M., Scheinman, R. and Frank, T. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 7826–7830.
26. Burke, D.T., Carle, G.P. and Olson, M. (1987) *Science*, **236**, 806–816.
27. Coulson, A., Sulston, J., Brenner, S. and Karn, J. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 7821–7825.
28. Coulson, A. and Sulston, J. (1988) In: *Genome Analysis: A Practical Approach*, (ed) K. Davies, pp.19–39.
29. Hauge, B. and Goodman, H., Personal communication.
30. Garza, D., Ajioka, J.W., Burke, D.T. and Hartl, D.L. (1989) *Science*, **246**, 641–646.
31. Hoheisel, J.D. and Lehrach, H., Personal communication.
32. Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) *Molecular Cloning, A Laboratory Manual*. Cold Spring Harbor.
33. Gibson, T.J., Rosenthal, A. and Waterston, R.H. (1987) *Gene*, **53**, 283–286.
34. Murray, N. (1983) In: *Lambda II*. (eds.) Hendrix, R.W., Roberts, J.W., Stahl, F.W. and Weisberg, R.A. Cold Spring Harbor.
35. Saunders, R.D.C., Glover, D.M., Ashburner, M., Sidén-Kiamos, I., Louis, C., Monastirioti, M., Savakis, C. and Kafatos, F.C. (1989) *Nucl. Acids Res.*, **17**, 9027–9037.
36. Feinberg, A.P. and Vogelstein, B. (1984) *Anal. Biochem.*, **137**, 266–267.
37. Sulston, J., Mallett, F., Staden, R., Durbin, R., Horsnell, T. and Coulson, A. (1988) *CABIOS*, **4**, 125–132.
38. Sulston, J., Mallett, F., Durbin, R. and Horsnell, T. (1989) *CABIOS*, **5**, 101–106.
39. Atherton, D. and Gall, J.G. (1972) *Drosophila Inf. Serv.*, **49**, 131–133.
40. Pardue, M.L. (1986) In: *Drosophila: A Practical Approach*. (ed.) D.B. Roberts. pp.111–137. IRL Press, Oxford.
41. Rasch, E.M., Barr, H.J. and Rasch, R.W. (1971) *Chromosoma*, **33**, 1–18.
42. Hinton, T. (1942) *Genetics*, **27**, 119–127.
43. Rudkin, G. (1982) *Results and Problems in Cell Differentiation*, **5**, 59–85.
44. Garza, D., Ajioka, J.W., Carulli, J.P., Jones, R.W., Johnson, D.H. and Hartl, D.L. (1989) *Nature*, **340**, 577–578.
45. Johnson, D. (1990) *Genomics*, **6**, 243–251.
46. Wesley, C.S., Ben, M., Kreitman, M., Hagig, N. and Eanes, W.F. (1990) *Nucl. Acids Res.*, **18**, 599–603.
47. Ludecke, H.-J., Senger, G., Claussen, U. and Horsthemke, B. (1989) *Nature*, **338**, 348–350.
48. Pirrotta, V. (1986) In: *Drosophila - A Practical Approach*. (ed.) D.B. Roberts. IRL Press Oxford, pp.83–110.
49. Dowsett, A.P. and Young, M.W. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 4570–4574.
50. Dowsett, A.P. (1983) *Chromosoma*, **88**, 104–108.
51. Martin, G., Wiernasz, D. and Schedl, P. (1983) *J. Mol. Evol.*, **19**, 203–213.
52. Cacccone, A., Amato, G.D. and Powell, J.R. (1988) *Genetics*, **118**, 671–683.
53. Werman, S.D., Davidson, E.H. and Britten, R.J. (1990) *J. Mol. Evol.*, **30**, 281–289.
54. Lemeunier, F. and Ashburner, M. (1976) *Proc. Roy. Soc. London*, **193B**, 275–294.
55. Knott, V., Rees, D.J.G., Cheng, Z. and Brownlee, G.G. (1988) *Nucl. Acids Res.*, **16**, 2601–2612.
56. Lander, E.S. and Waterman, M.S. (1988) *Genomics*, **2**, 231–239.
57. Bolshakov, V.N., Zharkikh, A.A. and Zhimulev, I.F. (1985) *Chromosoma*, **92**, 200–208.
58. Sulston, J., Personal communication.
59. Olson, M., Hood, L., Cantor, C. and Botstein, D. (1989) *Science*, **245**, 1434–1435.